Application of Regression Models in Convective Storm Nowcasting

Patrik Benáček[1,2], Michaela Valachová[1,2], Hana Kyznarová[3]

1. Central Forecasting Office, Czech Hydrometeorological Institute, Prague
2. Department of Atmospheric Physics, Faculty of Mathematics and Physics, Charles University
3. Radar Department, Czech Hydrometeorological Institute, Prague

Introduction

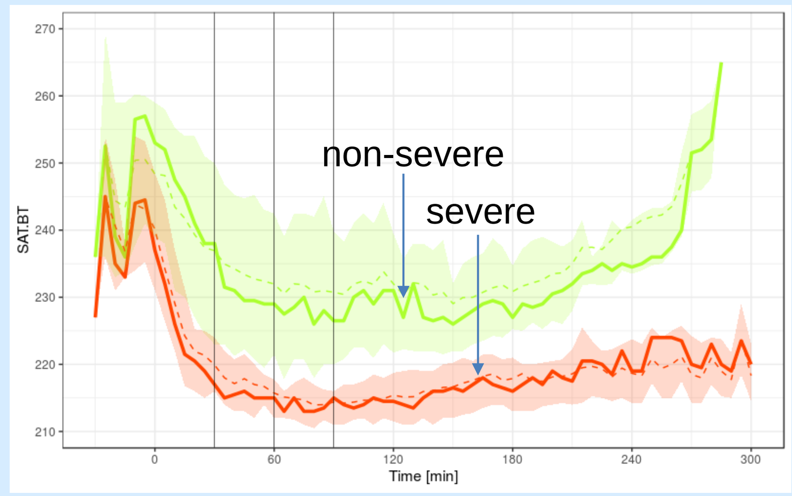
Nowcasting of convective storms is one of the most challenging tasks for operational weather forecasters. Remote sensing is a vital source of information about present storm occurrence and development in a high spatial and temporal resolution – mostly up to few kilometres every 5 min.

From the forecasters' experience, combining real-time data into only one or two products is desirable to supply credible and clear information for crucial decisions. Therefore, we employ regression models to obtain an objective information about the severe storm probability, having a potential to improve a real-time warning process.

Motivation

Is it going to be a severe storm?

- Already in the first 30 min of the storm lifetime, there is a significant difference in minimum brightness temperature in IR10.8 (MSG/SEVIRI).
- The storm lifetime starts when the radar reflectivity exceeds 30 dBz.



Data

Evolution of 72 isolated convective storms from 2016 and 2017, which formed in the region of Central Europe, is studied by means of multi-sensor observations from the first 30, 60 and 90 min of the monitored storm lifetime:

- response variable: severe/non-severe (24/48) classification based on the European Severe Weather Database (ESWD)
- independent variable (IV): total 54 IVs from radar (RAD), lightning (LSD) and satellite (SAT) observations

RAD	Description	Unit
AREA	RC area, 30 dBZ reflectivity threshold	km^2
VOL	RC volume, reflectivity above 4 dBZ	km^3
VOL44	RC volume, reflectivity above 44 dBZ	km^3
MAX_R	max of reflectivity within RC	dBZ
TOP x	max height of reflectivity x from -10 to 70 dBZ (5 dBZ step)	m a.s.l.
MAX_R_height	mean height of the area with the highest reflectivity of RC	m a.s.l.
VIL	Vertically Integrated Liquid (Greene and Clark, 1972)	kg/m^2
VILD	VIL density	g/m^3
VIL_SUM	VIL summary	kg/m^3
SHI	Severe Hail Index (Witt et al., 1998)	-
POH	Probability of Hail (Waldvogel et al., 1978)	%
POSH	Probability of Severe Hail (Witt et al., 1998)	%
MEHS	Maximum Estimated Hail Size (Witt et al., 1998)	mm

Tab. 1: RAD IVs derived by CELLTRACK (Kyznarová and Novák, 2009), based on the radar reflectivity and sounding data. For all the above variables, values are computed within the area of the reflectivity core (RC) together with their maxima, medians or means during the storm life-cycle.

SAT	Description	Unit
BT	Minimum brightness temperature (BT) pixel in IR10.8	K
dBT	Change of min BT in 5-min interval	K
Cooling15	Highest cooling rate of the min BT in 15-min interval	K

LSD	Description	Unit
LJ	Maximum lightning jump in the storm lifetime (Farnell et al., 2017)	-
nstroke	Number of all strokes together in 5-min interval	-
nstroke_x	As above, where x represents all CG/CG+/CG-/CC in 5-min interval	-
max_curr_0/3	Maximum peak current of CG/CC strokes	kA
sum_curr_0/3	Sum of peak current of CG/CC strokes	kA

Tab. 2: SAT IVs (top) from MSG/SEVIRI 5min Rapid Scan and LSD IVs (bottom) from CELDN, detected for the storm area by hierarchical clustering in R (www.r-project.org).

Variable selection

- **Multicollinearity:** strongly correlated IVs with VIF>4 are excluded from a full model. For RAD, the IVs are preselected based on scientific knowledge.
- **Sign OK:** Selection of IVs based on plausibility of a sign of the multivariable regression coefficient.
- **Stepwise backward**: elimination of IVs with p-value > 0.157, contributing more likely to noise than to the model predictive information.

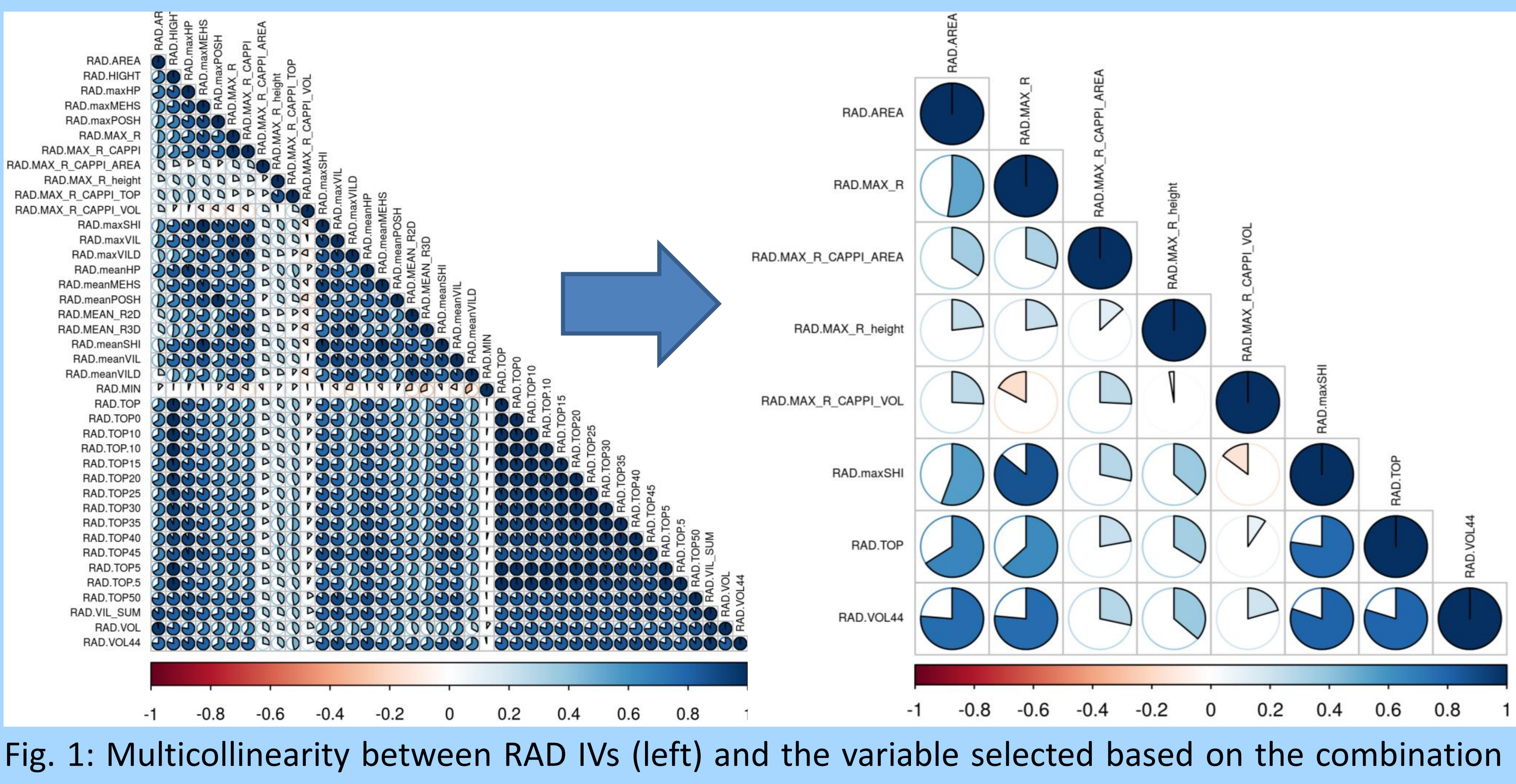


Fig. 1: Multicollinearity between RAD IVs (left) and the variable selected based on the combination of scientific knowledge and VIF<4 criteria (right).

Method	Description
Full model	Inclusion of all candidate predictors without the variable selection procedure.
Events per variable (EPV)	The number of cases relative to the number of IVs.
Variance inflation factor (VIF)	Identification of correlations between IVs in regression model. VIF of predictor x_j is calculated based on the linear relationship between the predictor x_j and the other independent variables $[x_1, x_2, \ldots, x_j] \setminus \{x_j\}$.
Stepwise backward elimination	Repeat: remove the most insignificant IVs and reestimate the model. Stop: no insignificant IV is left.
SignOK	Repeat: remove IV, if its multivariable sign is different from the univariable sign. Stop: all selected IVs have a correct sign.
Leave-one-out cross-validation (LOOCV)	Repeats over sample size N: take N-1 cases to build the model and test results against the remaining single case.
Odds ratio (OR)	The mean change in response variable for one unit of change in IV, while holding other IVs in the model constant.
Recall (RE), Precision (PR), F1-score	Evaluation of classification models; RE: x% severe storms that are correctly classified. PR: x% of classified severe storms is really severe. F1 score: a combination of PR and RE.

Building model

Regression modelling in the small data set (EPV<10) has a problem with convergence of the maximum likelihood formula. Therefore, after the variable selection, we employ the Logistic Regression Model (LRM) individually for each RAD, LSD, and SAT observations (Fig. 2).

	Dependent variable: sev.phenomena		
	Radar (1)	Lightning (2)	Satellite (3)
RAD.AREA	1.010 p = 0.013**		
RAD.TOP	1.000 p = 0.157		
LSD.LJ		2.091 p = 0.008***	
SAT.BT			0.894 p = 0.0005***
Observations	63	63	63
Log Likelihood	−29.570	−33.773	−29.744
Akaike Inf. Crit.	65.140	71.545	63.487
Note:			*p<0.1; **p<0.05; ***p<0.01

Fig. 2: The LRM IVs and odd ratios estimated separately for RAD, LSD and SAT observations. The training data are from the first 30 min of the storm lifetime.

The regularized regression model Elastic Net (ENet) is employed for a full model involving two regularization terms α and λ in the maximum likelihood formula. The ENet allows: i) selection of relevant IVs, and ii) shrinking regression coefficients towards zero so that we avoid overfitting of regression coefficients. The optimal penalty terms are searched by varying α over a grid of 0-1 by 0.02, searching for the best performance of each λ by the LOOCV (Fig. 3).

Predictors	Coefficient	Odd_ratio
(Intercept)	5.995	401.259
RAD.AREA	0.005	1.005
LSD.sum_curr_0_pos	-0.005	0.995
LSD.LJ	0.218	1.244
SAT.BT	-0.033	0.967

Fig. 3: The LRM IVs and odd ratios estimated separately for RAD, LSD and SAT observations. The training data are from the first 30 min of the storm lifetime.

	Recall [%]	Precision [%]	F1 Score
LRM-RAD	64	78	0.70
LRM-LSD	36	89	0.52
LRM-SAT	68	62	0.65
ENet	75	94	0.83

Fig. 4: Validation of the LRMs and ENet models based on the LOOCV method.

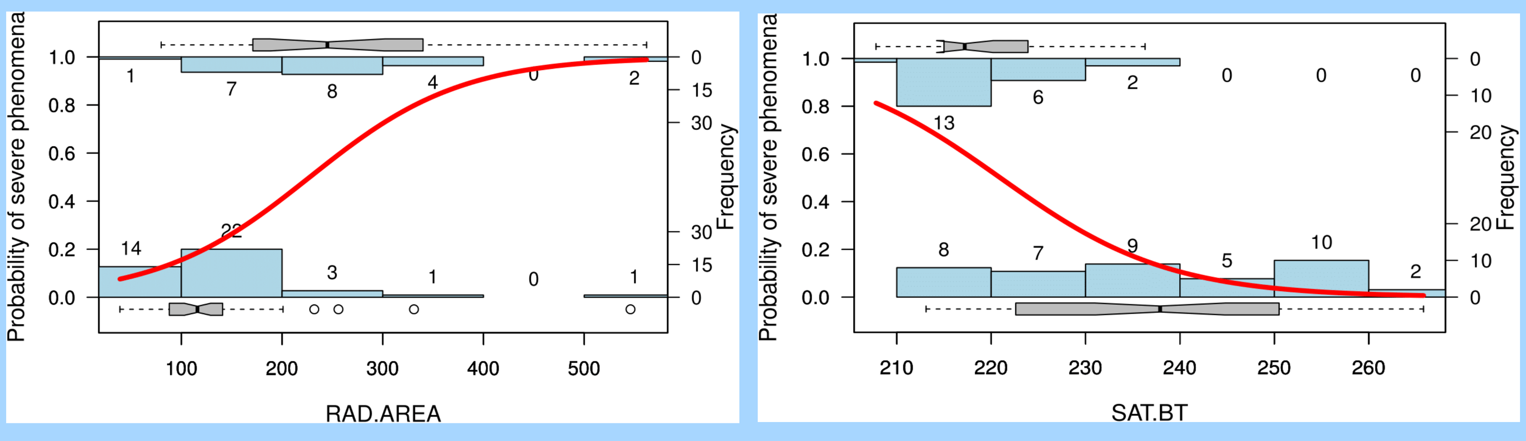


Fig. 5: Plot of the probability of the storm severity (0/1) vs. RAD.AREA (left) and SAT.BT (right), together with conditional distributions. Data are from the first 30 min of the storm lifetime.

Conclusion

- Based on regression models, the essential severe storm predictors in the first 30 min of the storm lifetime are:
  - maximum lightning jump,
  - radar echo-tops and area of identified reflectivity cores,
  - minimum BTs in the 10.8 µm spectral band.
- The elastic net:
  - handled the model stability problem in the small data set,
  - selected objectively similar severe storm predictors,
  - increased significantly the model performance.

Acknowledgement and references

We acknowledge CHMI, EUMETSAT and SIEMENS AG for providing data used in this study. ESSL and all contributors to ESWD deserve special thanks for their valuable effort in collecting information about severe weather occurrence in Europe.

Part of this work was supported by Ministry of the Interior of the Czech Republic, project "Program bezpečnostního výzkumu pro potřeby státu 2016 - 2021 (BV III/2 - VZ)" [VH20172020017].